ESTIMATION OF AUDITORY SPATIAL CUES FOR BINAURAL CUE CODING

Frank Baumgarte and Christof Faller

Media Signal Processing Research, Agere Systems, Murray Hill, NJ, U.S.A.

ABSTRACT

Binaural Cue Coding (BCC) offers a compact parametric representation of auditory spatial information. This representation can be applied to stereophonic or multi-channel audio compression. It allows to reconstruct the spatial image given a mono audio signal and spatial cues that require only a low data volume. This paper focuses on the extraction of the spatial cues from a stereophonic signal with a BCC analyzer. Results of a first subjective quality evaluation confirm that this technique is able to approximate the spatial image of critical reference signals. Thus BCC has the potential to offer highly efficient stereophonic or multi-channel coding, even at very low bit rates that currently permit coding of only one mono channel.

1. INTRODUCTION

The data rate of traditional stereophonic or multi-channel sub-band audio coding schemes like MPEG-2 AAC [1] scales with the number of channels. Compared to a proportional growth of bit rate for independent stereo coding, joint coding techniques, such as "M/S stereo", "intensity stereo" [1] and "inter-channel prediction" [2] can achieve a smaller growth with number of channels. However, the resulting bit rate is always considerably higher than needed for the corresponding mono signal. If the target bit rate falls below a certain threshold, the tradeoff between audio bandwidth, coding artifacts, and number of channels for quality optimization typically dictates to use only one mono channel.

To overcome the problem of the approximately linear increasing data rate with the number of audio channels, binaural cue coding (BCC) [3, 4] can be applied. BCC adds only a small fraction to the mono channel data rate to enable a stereophonic or multi-channel reconstruction. Figure 1 shows an example of the proposed stereo coding scheme, that transmits one compressed mono audio signal along with BCC side information. The mono signal is generated by adding the input channels.

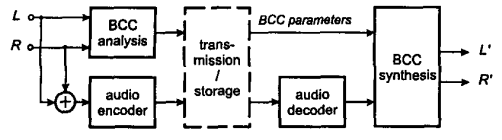

Fig. 1. BCC-based audio coding scheme.

The BCC analyzer extracts the perceptually most relevant binaural spatial cues from the stereophonic or multi-channel input signal. The BCC synthesizer reconstructs a stereophonic or multi-channel signal by inserting the cues into the decompressed mono signal. It aims to achieve the same spatial auditory image at the receiver as would appear from the analyzer input signal. The extraction of binaural cues does not involve sound source separation since the spatial cues can be identified without explicit knowledge of the sound sources. In fact, the cues can be extracted from an arbitrary complex mixture of sound sources.

The perceptually relevant binaural spatial cues are well known from psychoacoustics, especially for single sound sources. Binaural models are available that quantitatively predict a wide range of psychoacoustic data for binaural detection and localization [5]. Since these models are usually verified for a single source only, a generalization is necessary for the binaural cue extraction from arbitrary audio signals.

The most important parameters are the interaural time difference (ITD) and the interaural level difference (ILD). Since in general binaural signals are different from loudspeaker channels, we will use the terms inter-channel time and level differences (ICTD and ICLD) to characterize the spatial cues contained in the loudspeaker channels. The BCC analyzer estimates the ICLD and ICTD in each "critical" band associated with the effective spectral resolution in binaural hearing. The BCC synthesizer inserts the cues into the mono signal by sub-band amplitude or time-delay panning.

This paper is focussed on the analysis of stereophonic audio signals for BCC. the interested reader is referred to [3] for details about the insertion of spatial cues into a mono signal. A first implementation for extracting the ICTDs and ICLDs is described in Section 2. The subjective assessment procedure using the BCC reconstructed stereophonic signal, the unmodified reference, and "anchor" signals is given in Section 3. Results are given in Section 4. In Section 5 some conclusions are drawn.

2. ESTIMATION OF AUDITORY SPATIAL CUES

BCC takes advantage of the auditory system's limited precision in sound source localization. For single sources many of these limitations were psychoacoustically measured and modeled in the past [5]. From psychoacoustics it is known that the direction of the auditory event in the horizontal plane (azimuth) depends mainly on the ILD and ITD created by the sound event at the two eardrums [6]. Successful models for predicting the direction or the detectability of direction changes are generally based on a pre-processing scheme closely related to the presumed function of the auditory system. The pre-processing frequently includes a spectral decomposition into critical bands and an envelope generation for bands above a center frequency of about 1.5 kHz. Many models are based on a subsequent cross-correlation analysis of these signals of the left and right side.

Subject of this paper is an implementation of a BCC analyzer that employs the Cochlear Filter Bank (CFB) [7, 8] to obtain a critical band representation of each audio channel as shown in Fig. 2.

0-7803-7402-9/02/$17.00 ©2002 IEEE

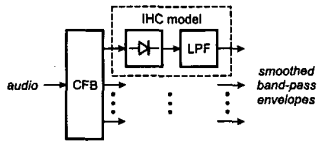

Fig. 2. Block diagram of Cochlear Filter Bank (CFB) and inner hair cell (IHC) model.

The band-pass outputs of the CFB are half-wave rectified and low-pass filtered. The low-pass filter (LPF) is composed of two identical cascaded first order filters with a cutoff frequency of $f_{c,\mathrm{IHC}}$ according to (1). The center frequency of the CFB band in Hz is denoted by f_{CFB} and given in [8]. The parameter f_0 is chosen to be $f_0 = 300$ Hz.

$$f_{c,\mathrm{IHC}} = \begin{cases} f_0 & \text{if } f_{\mathrm{CFB}} < f_0 \\ f_0 \left(\frac{f_{\mathrm{CFB}}}{f_0}\right)^{0.25} & \text{if } f_{\mathrm{CFB}} \geq f_0 \end{cases} \quad (1)$$

A smooth envelope is derived at the outputs of the CFB bands at medium and high center frequencies. For simplicity all outputs are referred to as band-pass envelopes even though at low center frequencies the waveform is still present. The CFB outputs have a maximum delay of 10 ms that decreases with increasing center frequency. The delay of all bands is equalized by adding the necessary delay in each band. This simplifies the application of the estimated cues to a BCC synthesizer that is based on a uniform filter bank or transform with constant delay.

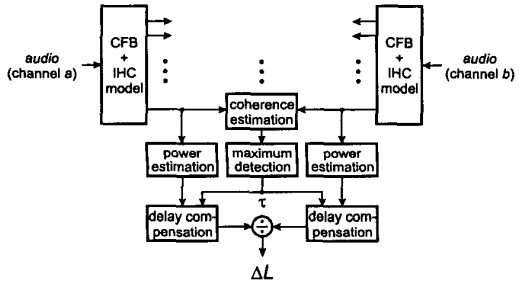

Fig. 3. Block diagram of BCC analyzer.

Figure 3 shows the derivation of ICTDs, τ, and the ICLDs, ΔL, in each CFB band for a pair of channels (e.g. left and right channel of a stereophonic signal). The estimation of the ICTDs is based on the coherence function in each band which is a normalized cross-correlation measure. The coherence function estimate $\hat{\gamma}_{xy}$ is derived from the cross-correlation estimate $\hat{\phi}_{xy}$ normalized by the auto-correlation estimates $\hat{\phi}_{xx}$ and $\hat{\phi}_{yy}$ of the signals in both channels a and b according to (2) for $m \geq 0$ and (3) for $m < 0$. The time shift m is expressed as number of sampling intervals.

$$\hat{\gamma}_{xy}(m,i) = \hat{\phi}_{xy}(m,i) \left[\hat{\phi}_{xx}(0,i-m)\, \hat{\phi}_{yy}(0,i)\right]^{-0.5} \quad (2)$$

$$\hat{\gamma}_{xy}(m,i) = \hat{\phi}_{xy}(m,i) \left[\hat{\phi}_{xx}(0,i)\, \hat{\phi}_{yy}(0,i+m)\right]^{-0.5} \quad (3)$$

The smoothed band-pass envelopes with removed DC component in one CFB band of the two input audio channels a and b are denoted x and y, respectively. The index of the current sampling interval (time index) is i. The time shift in sampling intervals between the envelopes x and y is m. The cross-correlation function is estimated recursively using (4) for $m \geq 0$ and (5) for $m < 0$.

$$\hat{\phi}_{xy}(m,i) = w\, \hat{\phi}_{xy}(m,i-1) + [1-w]\, x(i-m)\, y(i) \quad (4)$$

$$\hat{\phi}_{xy}(m,i) = w\, \hat{\phi}_{xy}(m,i-1) + [1-w]\, x(i)\, y(i+m) \quad (5)$$

The time constant of the exponential estimation window is determined by w. It was adjusted such that the estimated time and level differences based on the coherence are able to follow changes in the input correlation fast enough while maintaining a reasonable stable result for a stationary ICLD and ICTD for natural sound sources like speech or vocals. A good compromise is achieved with $w = 0.998$. If (4) and (5) are interpreted as recursive low-pass filtering with the filter coefficient w the corresponding cutoff frequency is about 10 Hz for a sampling rate of $f_s = 32$ kHz.

The auto-correlation estimates in (2) and (3) used for the normalization are estimated according to (6) and (7). The same factor w as in (4) and (5) must be used here to maintain the desired coherence range between -1 and 1.

$$\hat{\phi}_{xx}(0,i-m) = w\, \hat{\phi}_{xx}(0,i-m-1) + [1-w]\, x^2(i-m) \quad (6)$$

$$\hat{\phi}_{yy}(0,i+m) = w\, \hat{\phi}_{yy}(0,i+m-1) + [1-w]\, y^2(i+m) \quad (7)$$

The ICTD is estimated by locating the maximum of the coherence function $\hat{\gamma}_{xy}(m,i)$ with respect to m. If the maximum is located at $m = m_{max}$, the ICTD is $\tau = m_{max}/f_s$ seconds. The ICLD estimation is based on the ratio of the estimated band powers. The power estimation uses a recursive low-pass filter applied to the squared inner hair cell model outputs. The filter cutoff frequency is about 50 Hz. The ICLD, ΔL, is the ratio of the delay compensated power estimates from both channels and converted to the logarithmic (dB) domain.

The coherence function is computed for a limited symmetrical range of delays m with respect to zero delay because auditory localization based on ITDs "saturates" at the extreme left or right side for delays larger than approximately 1 ms. However, we currently use a delay range of ± 1.6 ms to get an improved ICLD estimate if larger ICTDs are present.

2.1. Results of cue estimation

In the following the performance of the ICTD and ICLD estimation is demonstrated for speech signals. For that purpose different stereo signals were created by amplitude and/or time-delay panning and superposition of two separate talkers. The estimated inter-channel cues are compared to the "ideal" cues applied. Figure 4 shows the estimated power of the two separate one-channel talker signals at a center frequency of 1008 Hz. The panning of these two mono signals was done according to Table 1 to create a stereophonic signal with ICLDs only (A), with ICTDs only (B), and a combination of both (C). The estimated ICTDs in Fig. 5 for B and C show an instantaneous transition between the ICTDs of both talkers since the coherence function can have two maxima corresponding to the two delays. Since the larger maximum is chosen for the ICTD estimation, basically a switching between both values occurs. The ICTD estimate for A is almost ideally zero except for a few single values. Figure 6 shows the estimated ICLDs for all three signals. It appears almost identical for A and C as expected. Due to the overlap of both talkers the ICLD gradually changes between the ICLD of one talker to the ICLD of the other. For B the estimate is close to the applied ICLD of zero as desired.

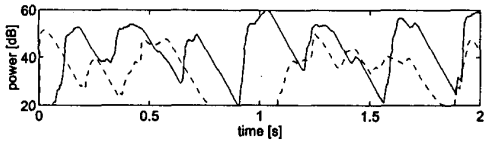

Fig. 4. Estimated signal power over time of the two talkers in the band centered at 1008 Hz. Talker 1: dashed, talker 2: solid.

	talker 1		talker 2	
	ICTD [dB]	ICLD [ms]	ICTD [dB]	ICLD [ms]
A	0	10	0	−10
B	0.6	0	−0.6	0
C	0.6	10	−0.6	−10

Table 1. Parameters of synthesized stereophonic signals A, B, C.

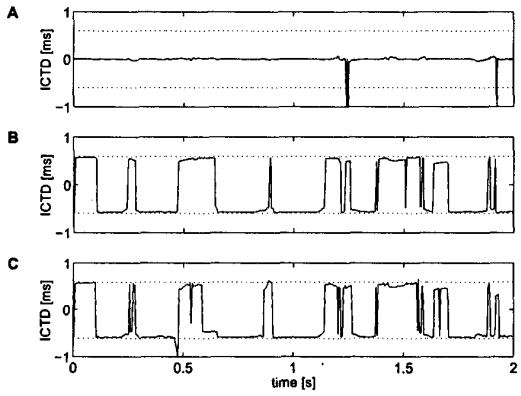

Fig. 5. Estimated inter-channel time differences (ICTD) for the three synthesized stereophonic signals according to Table 1 in the band centered at 1008 Hz.

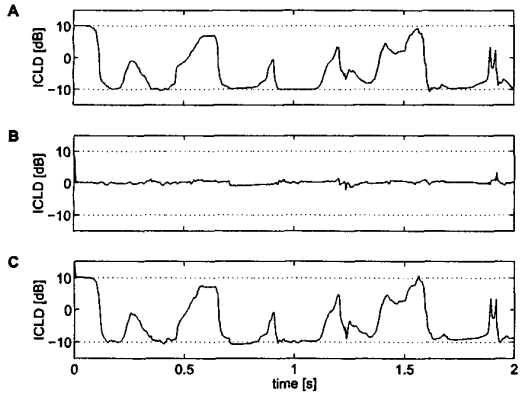

Fig. 6. Estimated inter-channel level differences (ICLD) for three stereophonic signals according to Table 1. Shown for the band centered at 1008 Hz.

3. QUALITY ASSESSMENT

In a first attempt to assess the quality and the reproduced spatial image of BCC we used synthesized stereophonic signals with two or three discrete phantom sources. The phantom sources were created by amplitude panning which imposes an ICLD, $\Delta L_{\mathrm{ref},n}$, onto the time-aligned signals of the n-th source in the two stereo channels. The overall level is adjusted such that the power of the synthesized stereo signal is equal to the sum of the source powers.

The choice of synthesized signals as opposed to natural stereo recordings is motivated by having better control over the spatial image and strictly defined parameters for creating the image. A further advantage of the synthesized signals is the lack of spatial reverberation and reflections that add more perceptual dimensions to be covered in a subjective assessment.

Using only ICLDs for loudspeaker playback is motivated by the dominant role of ITDs at low frequencies for auditory localization. In that frequency range ITDs are chiefly determined by ICLDs due the effect of the listener's head on the sound field [6]. However, for headphone playback it is necessary to supply the ICTDs at low frequencies to achieve a good spatial reproduction. This was demonstrated in unpublished BCC experiments and is in line with the literature [6].

Four different categories of signal sources were used: single talkers (D), solo vocals (E), keyboard instruments (F), and percussive instruments (G). Four reference signals (D2...G2) were generated by mixing two sources of the same category with ICLDs of 10 and −10 dB. Another four reference signals (D3...G3) were generated by mixing three sources of the same category with ICLDs of 10, −10 dB, and 0 dB. Sources of the same category were mixed because they are most likely to have an impact on each others spatial image due to their similar time-frequency characteristics.

To facilitate the evaluation of the listening test results, two types of anchor signals were also presented in the test. For the first type the ICLDs are sinusoidally modulated over time with a frequency of 0.5 Hz to create moving phantom sources. The ICLD varies between 10 and 5 dB instead of 10 dB in the reference, between −10 and −5 dB instead of −10 dB, and between −2.5 and 2.5 dB instead of 0 dB. The second type adds localization blur by modifying the ICLDs of every other critical band. The modified bands have 5 dB ICLD instead of 10 dB in the reference and −5 dB ICLD instead of −10 dB. The ICLD of 0 dB in the reference is replaced by alternating the ICLDs in the bands between −2.5 and 2.5 dB.

The BCC processed signals were obtained by analyzing the stereophonic input (reference) signal to generate the spatial cues and by creating the mono signal. For an audio bandwidth of 16 kHz, ICLDs of 21 critical bands were estimated and transmitted to the BCC synthesizer every 128 samples (4 ms) without quantization. The BCC stereophonic signal was reconstructed from the mono signal by inserting the generated spatial cues with a synthesizer as described in [3]. The synthesizer was operated with a size 512 FFT and a time-window length of 256.

4. RESULTS

Nine trained subjects were asked to assess the overall quality of the anchor signals and the BCC processed signals with respect to the reference in a listening test over loudspeakers. The subjective test was based on ITU-R BS.1116 [9] using the ITU-R five-grade-impairment scale. Additionally, subjects were asked to specify

the kind of degradation they perceive. They were given a choice of specifying "reduced image width", "reduced image stability", "increased image blur", or "other degradations".

Some trends can be observed from the results in Fig. 7. The quality of the anchor signals is almost equal among the two-phantom-source signals and among the three-phantom-source signals that have a consistently higher perceived quality. However, the average quality of BCC appears not to depend on the number of phantom sources but it depends considerably on the signal category. Other degradations, such as sound coloration or other artifacts, were insignificant.

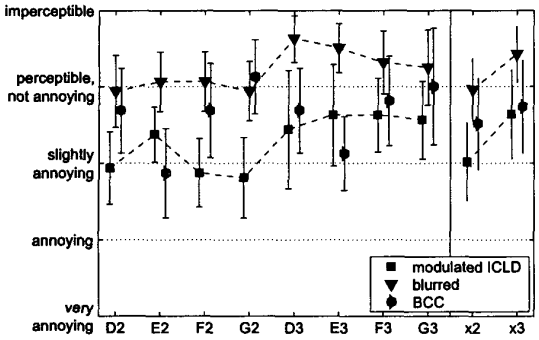

Fig. 7. Impairment gradings according to the ITU-R five-grade scale and 95%-confidence intervals. (Averaged gradings over two-source mixes (x2) and three-source mixes (x3)).

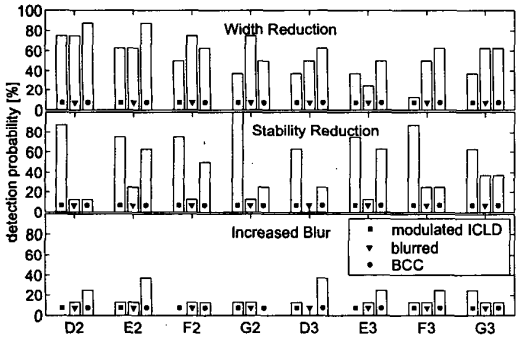

Fig. 8. Probability for detecting image width reduction, stability reduction, and increased blur.

Figure 8 shows that the modulated ICLD anchor is detected with moderate probability of having an image-width reduction and high probability of having reduced stability. The spectrally blurred anchor was detected most of the time by a reduced image width. Its increased blur was only detected by one subject on average for most items. Subjects apparently show less sensitivity to these spectral modification than to the ICLD modulation. In comparison, BCC has an almost constant moderate probability for image width reduction. However, image stability and blur depend on the signal category. Stability reduction is less probable for BCC than for the modulated anchor. Image blur of BCC is slightly larger than for the blurred anchor.

These results suggest, that BCC is able to reconstruct two or three phantom sources in a stereo signal from mono with a signal-dependent degradation of the spatial image. Since other artifacts are not significant, we argue that this degradation is more tolerable than typical coding artifacts of stereo coders at the same low target bit rate (see [4] for details).

5. CONCLUSIONS

The presented BCC analyzer is based on binaural models to extract auditory spatial cues from a stereophonic signal. In a first experiment only ICLDs were incorporated to reconstruct the spatial image. Subjective test results from critical material reflect the currently achieved overall quality and the kind of degradation.

The quality of BCC is determined by a degradation of the auditory spatial image only. The degradation depends on the audio material. However, it does not change if the number of phantom sources is increased from 2 to 3. Based on our experience, the spatial degradation introduced by BCC is less annoying than typical coding distortions created by traditional stereo coders at bit rates in the order of 64 kbit/s and below [4]. This suggests that BCC enables higher quality stereo and multi-channel coding at these bit rates than was possible before. Future tuning of the time-frequency resolution and other factors is expected to enhance the quality.

6. ACKNOWLEDGEMENTS

We thank Aki Härmä, Peter Kroon, and Jens Meyer for valuable comments on an earlier draft of this paper.

7. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11 "Coding of moving pictures and audio – MPEG-2 Advanced Audio Coding". *ISO/IEC 13818-7 International Standard*, 1997.

[2] Fuchs, H. "Improving joint stereo audio coding by adaptive inter-channel prediction" *Proc. 1993 IEEE WASPAA*, New Paltz, NY, Oct. 1993.

[3] Faller, C. and Baumgarte, F. "Efficient representation of spatial audio using perceptual parametrization," *Proc. 2001 IEEE WASPAA*, New Paltz, NY, Oct. 2001.

[4] Faller, C. and Baumgarte, F. "Binaural cue coding: A novel and efficient representation of spatial audio)," *Proc. ICASSP 2002* (submitted), Orlando, Florida, May 2002.

[5] Stern, M.S. and Trahiotis, C. "Models of binaural perception". In Gilkey, H. R. and Anderson T. R., *Binaural and spatial hearing in real and virtual environments*," Erlbaum, 1997.

[6] Blauert, J. *Spatial Hearing: The psychophysics of human sound localization*. MIT Press, 1996.

[7] Baumgarte, F. "A computationally efficient cochlear filter bank for perceptual audio coding," *Proc. ICASSP 2001*, Salt Lake City, May 2001, pp. 3265–3268.

[8] Baumgarte, F. "A psychoacoustic model for audio coding based on a cochlear filter bank," *Proc. 2001 IEEE WASPAA*, New Paltz, NY, Oct. 2001.

[9] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *Rec. ITU-R BS.1116-1*, 1997.